CLAIMS

What is claimed is:

1. A method to process a document, comprising:

partitioning document text into a plurality of sentences;

for each sentence, assigning corresponding associated parts of speech to words, where assigning comprises applying a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments, to combine recognized chemical name fragments into a complete chemical name, and to assign the complete chemical name with one part of speech; and

parsing the sentence into its component parts based at least in part on the assigned parts of speech.

2. A method as in claim 1, where the complete chemical name is assigned a noun phrase part of speech.

3. A method as in claim 1, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

4. A method as in claim 1, where said plurality of dictionaries comprise a dictionary of stop words to

eliminate erroneous chemical name fragments.

5. A method as in claim 1, further comprising filtering recognized chemical name fragments using a list of stop words to eliminate erroneous chemical name fragments.

6. A method as in claim 1, where chemical name fragments are further recognized by using common chemical word endings.

7. A method as in claim 1, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

8. A method as in claim 1, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

9. A method as in claim 8, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

10. A method as in claim 8, where the characters comprise at least one of upper case C, O, R, N and H.

11. A method as in claim 8, where the characters comprise strings of at least one of lower case xy,

ene, ine, yl, ane and oic.

12. A method as in claim 1, comprising an initial step of tokenizing the document to provide a sequence of tokens.

13. A system for processing a text document, comprising:

a first unit for partitioning document text into a plurality of sentences;

a second unit, operable for each sentence, for assigning corresponding associated parts of speech to words, said second unit comprising sub-units to apply a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments, to combine recognized chemical name fragments into a complete chemical name, and to assign the complete chemical name with one part of speech; and

a third unit for parsing sentences into component parts based at least in part on the assigned parts of speech.

14. A system as in claim 13, where the complete chemical name is assigned a noun phrase part of speech.

15. A system as in claim 13, where said plurality of dictionaries comprise a dictionary of common

chemical prefixes and a dictionary of common chemical suffixes.

16. A system as in claim 13, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

17. A system as in claim 13, where said second unit further comprises a sub-unit for filtering recognized chemical name fragments using a list of stop words to eliminate erroneous chemical name fragments.

18. A system as in claim 13, where chemical name fragments are further recognized by using common chemical word endings.

19. A system as in claim 13, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

20. A system as in claim 13, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

21. A system as in claim 20, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

22. A system as in claim 20, where the characters comprise at least one of upper case C, O, R, N and H.

23. A system as in claim 20, where the characters comprise strings of at least one of lower case xy, ene, ine, yl, ane and oic.

24. A system as in claim 13, further comprising a tokenizer for tokenizing the document to provide a sequence of tokens.

25. A computer program product for storing in a computer readable form a set of computer program instructions for directing at least one computer to process a text document, comprising instructions for assigning corresponding associated parts of speech to words found in the document, where said instructions for assigning comprise instructions to apply a plurality of regular expressions, rules and a plurality of dictionaries to recognize organic chemical name fragments, to combine recognized organic chemical name fragments into a complete organic chemical name, and to assign the complete organic chemical name with one part of speech.

26. A computer program product as in claim 25, where the complete organic chemical name is assigned a noun phrase part of speech.

27. A computer program product as in claim 25, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

28. A computer program product as in claim 25, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

29. A computer program product as in claim 25, further comprising instructions for filtering recognized organic chemical name fragments using a list of stop words to eliminate erroneous fragments.

30. A computer program product as in claim 25, where chemical name fragments are further recognized by using common chemical word endings.

31. A computer program product as in claim 25, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between organic chemical name fragments as a function of context.

32. A computer program product as in claim 25, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

33. A computer program product as in claim 32, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon, where the characters comprise at least one of upper case C, O, R, N and H, and further comprise strings of at least one of lower case xy,

ene, ine, yl, ane and oic.

34. A computer program product as in claim 25, where said instructions for assigning operate on a sequence of tokens derived from document text.

35. A system comprising a plurality of computers at least two of which are coupled together through a data communications network, said system comprising a first unit for partitioning document text into a plurality of sentences; a second unit, operable for each sentence, for assigning corresponding associated parts of speech to words, said second unit comprising sub-units to apply a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments, to combine recognized chemical name fragments into a complete chemical name, and to assign the complete chemical name with one part of speech; and a third unit for parsing sentences into component parts based at least in part on the assigned parts of speech.

36. A system as in claim 35, where the complete chemical name is assigned a noun phrase part of speech.

37. A system as in claim 35, where a user of the system accesses the system through a data communications network.